网络信息资源保存的协作网络研究*

□ 吴振新 张智雄 王婷 / 中国科学院国家科学图书馆 北京 100190

摘要：协作保存网络指的是利用网络及相关工具软件为分布在不同地理位置的人们提供协同服务以进行保存工作和存档信息共享。文章从技术的角度，通过几个典型的Web archive协作保存网络案例，对如何构建协作保存网络来共享保存系统和资源进行了初步的研究和分析。该文为2009年第七期“网络信息资源保存”专题文章之一。

关键词：网络信息，长期保存，协作网络
DOI：10.3772/j.issn.1673—2286.2009.07.002

1 前言

随着网络信息资源保存（Web archive，简称WA）活动的不断发展，保存规模不断增大，对于保存活动的长期性和稳定性要求愈发强烈，协调和调度足够的社会资源，共同分担保存的风险和责任，合作进行保存活动已经成为保存机构的必然选择。

从目前国际上网络信息资源保存的合作情况看，合作内容覆盖了网络信息资源保存活动的全部过程，涉及政策、法律、经济、技术和管理等方方面面，并呈现出合作的多态性特点。在技术领域的合作中，从最初的标准规范制定，到保存工具的研发，随后是保存系统的建设，目前已经发展为构建虚拟协作保存网络，呈现了合作复杂化的发展趋势。

本文从技术的角度，对目前的WA项目中如何构建协作保存网络来共享保存系统和资源进行了初步的研究和分析。

2 网络信息资源的协作保存网络研究

协作网络指的是利用网络及相关工具软件为分布在不同地理位置的人们提供了信息共享和协同工作的平台环境。协作网络通常包含两种不同的工作模式：一种是指通过网络将大的任务或项目拆分为许多小的工作单元，使人们可以协同工作，提高工作效率；另一种方式是提供协同工作空间，使得一个工作团体可以轻松的共享资源，同时这些空间内的内容可以与网络上的其它资源实现无缝整合、透明连通。

协作保存网络是基于保存系统而构建的协作网络。国际网络保存联盟（International Internet Preservation Consortium，简称IIPC[1]）提出了基于OAIS[2]的Web archive系统技术体系框架（见图1），该框架覆盖了WA工作链中的所有过程，包括采集（harvest）、存储（Storage）、访问（Access）和索引与检索（Index &Search）等主要功能。从目前的WA项目看，基于WA工作链中的任何一个环节进行合作都可形成一个协作网络，而且不同环节间构建的协作网络因需求的不同也采用了不同的工作模式（包含混合的工作模式），下面将选取典型的案例对几种WA协作保存网络进行详细阐述。

* 本文系国家社会科学基金项目“网络信息资源保存的理论与方法研究”（项目编号：06BTQ025）的研究成果之一。

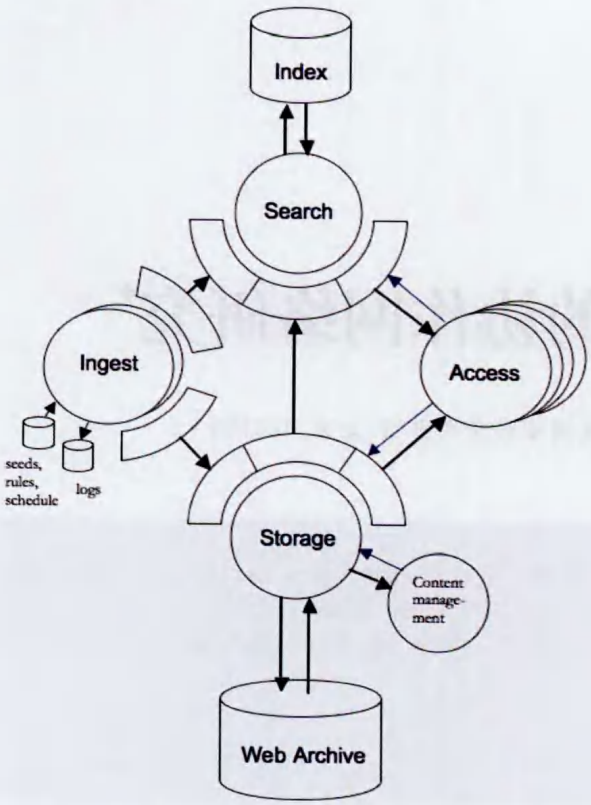

图 1　IIPC的WA系统技术体系框架[3]

2.1 基于采集的协作保存网络

所谓基于采集的协作是指网络信息资源保存工作在采集层面进行的协作。网络信息资源保存通过持续采集网络资源达到对不断变化的网络进行保存的目的，网络信息采集是WA的起点和基础，是长期保存的关键环节。这种协作方式的典型代表是澳大利亚PANDORA项目，它所采用的是任务分解的工作模式。

PANDORA项目[4]的合作伙伴分布在澳大利亚的各个州。PANDORA开发了一个综合的、基于网络的、可以远程工作的数字存档系统PANDAS。PANDAS的工作流程包括[5]：识别、选择主题；登记备选主题（描述元数据）；征求并记录存档许可；设置采集机制；实施采集；实施质量审核；存档；发布（包括为归档的资源组织用于发现、访问、呈现的相关元数据）。PANDAS调用HTTrack从网上采集资源，它也提供其他功能来调度采集过程。同时PANDAS系统也支持上传功能，可以从本地上传新的资源（可以是单个文件、多个文件甚至是整个网站），例如一个站点无法从网上正常收割而出版商通过其他方式（如FTP、CD）提供文件时就需要使用本地上传功能，系统还可通过邮件来提供文件的上传。上传文件功能只能由授权的PANDAS用户使用。

PANDAS系统的服务器设立在澳大利亚国家图书馆，每个合作伙伴通过国家图书馆分配的用户名和密码远程登录到系统进行操作。每个合作伙伴（作为集合管理员）按照各自制定的采集指南负责采集不同类型的资源，在选择采集的主题后通过PANDAS按照采集流程完成采集任务。

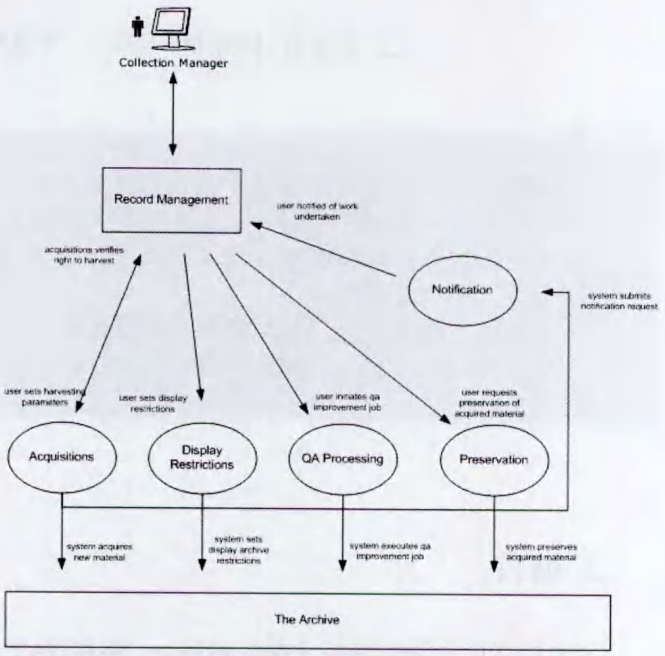

图2 PANDAS系统集合管理员工作流程[6]

统计表明，采集来的网络资源如果不进行流程控制和质量审核处理，约40%的资源无法保证使用。但对于长期保存而言，存档质量是非常重要的，从采集主题的选择到元数据编辑、数据采集、质量审核直至存档，都必须进行严格统一的管理，以确保存档资源的质量。对于海量的网络资源而言，这是一项需要巨大人工投入的工作，而PANDORA项目通过PANDAS系统将艰巨的采集任务进行了合理地分解，充分地调度各州图书馆的力量，并通过统一的流程控制和质量审查标准有效地保障了存档资源的质量，从而成为网络信息保存领域的一个成功范例。

英国UKWAC项目[7]也是看中PANDAS系统在分布式环境中的这一使用优势，因此选用PANDAS系统作为其网络信息资源保存系统构建了英国的网络信息协作保存网络。

这种基于采集的协作能够较好的保证数据的规范

和质量，合作者易于使用，同时由于合作伙伴可以自己制定采集指南，也使得其有相当的自主权，较适合有统一管理者的合作项目使用。

2.2 基于存储的协作保存网络

所谓基于存储的协作就是各个参与网络信息资源保存项目的机构在存储这一层面上进行协作，从而实现大容量的数据存储、快速的数据访问以及多数据备份。这种协作方式的代表项目是LC-SDSC的Chronopolis框架，这也是一个典型的通过网格实现的大规模存储的项目。

Chronopolis[8]联合了SDSC（圣地亚哥超级计算中心）、NCAR（美国国家大气研究中心）、UMIACS（马里兰大学高级计算机研究中心）三家机构，建立了一个具有三个节点的数据网格联盟。Chronopolis要求参与者至少需要在其网格节点上实现50TB的数字资源存储，并同时在这3个不同的物理点进行相互的数据备份。其结点图见图3 。

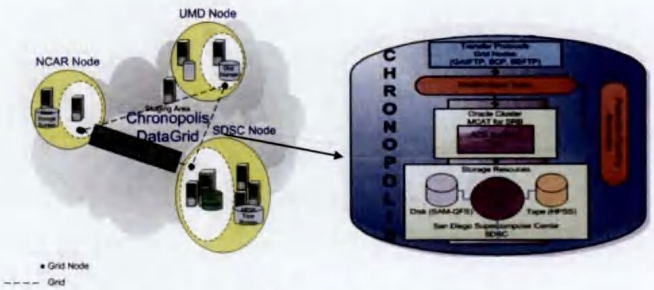

图3 Chronopolis的结点图[9]

Chronopolis项目采用了混合的协作模式，三个节点协同提供网格服务，同时又有不同的分工[8]：SDSC提供对所有数据的完整保存，构建网络服务与存储，提供SRB的支持服务，开发数据传输模块；NCAR 提供对所有数据的完整保存，提供存储和网络支持以及网络监测； UMIACS提供对所有数据的完整保存，提供高级数据服务，主要包括ACE（确保数字存档真实性的检测控制环境）、PAWN（支持长期保存的生产—存档工作流网络）、INCA（基于用户级的网格监测系统）。

如图4所示，三个节点间通过中转网格（Staging Grid）连接，中转网格负责接收来自提交方的数据，并对数据包进行完整性检查，出于安全的目的，中

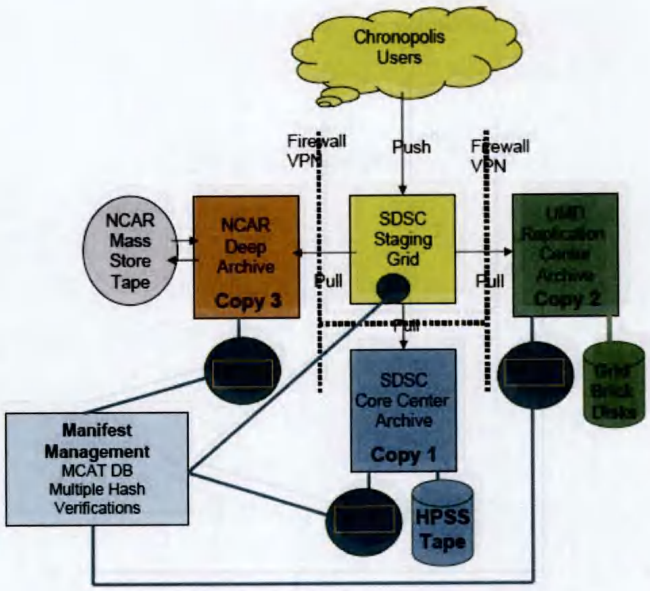

图4 Chronopolis工作流程图[9]

转网格的存储与其它存储是隔离的，通过检测的数据被独立地推送到三个存档结点的系统中，三个节点间通过GridFTP进行网格中大规模的并行数据传送。Chronopolis项目的网格是由SRB（SRB 是为在网格环境中实现数据管理而设计的一个分布式档案管理与资源存储系统，为使用者提供了一个存取系统、档案系统、数据库系统等多种异质存储系统的统一接口，涵盖了异质储存系统的特性）来实现的，在Manifest层提供额外的数据库管理的安全性保障和数据的完整性监测。每份数据资源都存有3个独立管理的副本，保证了数据的高可靠性和高可用性。

2006-2007年间，Chronopolis项目已经存储了国家虚拟气象台NVO总计3TB Hyperatlas格式的图片，国会图书馆PG图片收藏Prokudin-Gorskii（俄国摄影师）600G的图片，ICPSR总计2TB可获取的网络数据和NCAR总计3TB需要再重新分析的观测数据[8]。同时，随着项目的发展，节点的数量也可以继续增加，可以满足不断增长的存储容量的需求，它正在协商建立与其它网络（如NDIIPP network）的协作以构成更为广泛的协作网络。

作为一种协作的网格框架，Chronopolis节点之间的存储模块既相对独立又相互联合，从而使得数据具有高可用性和可靠性，网格的方式则融合了异构数据与异构系统。这种框架优秀的扩展性非常适合海量存储。另外，Chronopolis项目参与者彼此间不同的分工协作也优化了彼此的强项，增强了可靠性。但是，这种

协作模式需要解决结点之间的信任和安全问题，以及在结点之间大规模数据传输的问题。

2.3 基于访问的协作保存网络

基于访问的协作就是各参与项目在访问层面上进行协作，实现对分布存储的大容量数据的快速访问，从而最终达到资源共享的目的。Nordic Web Archive（NWA）项目[10]就是这种模式的典型代表。

NWA的目的是共同协作以建立起欧洲网络信息资源保存的合作机制，开发通用的技术和方法来支持各国的采集、存档，特别是访问，为此NWA开发了一系列网络信息资源保存的工具包。

在采集上，NWA的合作伙伴使用了Combine[11]、NEDLIB harvester[12]、HTTrack[13]、Heritrix[14]等采集器分别进行采集工作，在数据存储上也采用了不同的策略，包括不同的数据对象模型、不同的存储系统和媒介，其检索工具也各不相同，包括Excalibur[15]、Fast Search engine[16]、Lucene[17]等，它们的协作主要是体现在索引和访问模块上。如图5，NWA提供了一个Document Retriver模块，从Archive中提取存档的数字对象和相关元数据，按照NWA通用格式传输给Indexer模块，由Indexer建立所支持搜索引擎的索引。访问模块为用户提供搜索、浏览和导航的功能。当用户提交查询请求，访问模块调用搜寻引擎来寻找满足查询条件的对象；当用户请求特定网页文件，访问模块将调用Document Retriver模块返回存档的数据对象。

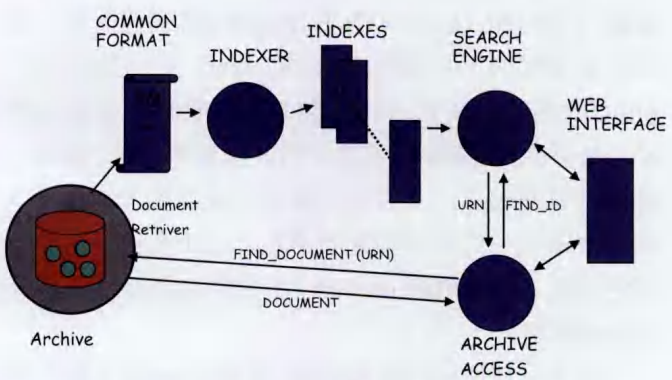

图5 NWA的技术架构及工作流程图[18]

在实际部署中，NWA顶层设置了一个适用于任何国家的超级发送结点，同时在挪威、瑞典、丹麦、冰岛、芬兰设置分布式发送子结点，每个子结点下再继续细分为n个索引结点（见图6）。为同步检索北欧保存的资源，国家发送结点需要作为前端的分布式结点，当用户进入访问界面提出检索请求，访问模块通过前端的国家发送结点向北欧所有的发布结点和索引结点发出请求，得到结果后再将结果返回给访问模块，访问模块再将此结果传递到浏览器供用户浏览。

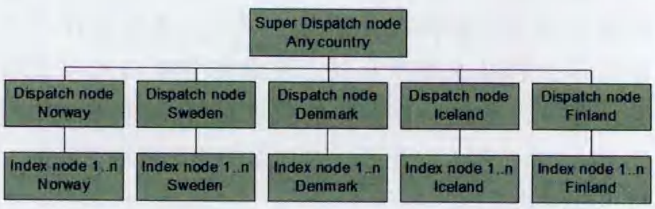

图6 NWA项目搜索引擎的架构[19]

这种多维度扩展的分布式检索架构可根据需求分别对查询结点（服务器）和索引结点（服务器）进行线性扩展，从而保障整体系统的可扩展性，同时在文本检索速度、执行的复杂度方面均有不错的表现。

NWA通过在索引和检索模块上的协作，使各国通过协作网络所提供的协同工作空间共享存档的资源，同时这些国家又保持了本国网络信息资源保存的独立性和特色（其采集的方式、采集策略的制定以及存储策略由各国自行制定），因此成为网络信息资源保存领域一个很有特点的协作案例。

实际上，基于访问的协作网络是最有效的资源共享模式，得到了很多项目的关注。目前还有一些项目通过其它技术方法也达到了同样的效果，葡萄牙的PWA（Portuguese Web Archive）项目[20]采用GAppA软件构建了协作访问网络，通过安装一个客户端应用程序，允许外部计算机加入计算机集群，该网络能够利用多个计算机集群联合对存档的数据提供远程访问[21]。美国的Web at risk项目[22]则通过Web资源集合注册（Web Collection Registry）来共享存档资源的元数据而为用户提供联合访问服务。

3 结语

网络信息资源保存的协作和基于网格的发展问题是一个新的、也是极具挑战性和实践性的研究课题。协作网络允许将大任务分割为多个子任务，而且结点间可以共享彼此的系统和资源，形成了虚拟的多保存系统以得到协作保存成效。这一问题的研究对于应对

日益增长的海量网络信息资源以及处于不同地域的保存系统之间通过协作实现资源最大化的共享具有重要的意义。同时在网络信息资源保存活动中，如何获得保存系统的鲁棒性与高效性是一个巨大的挑战，协作网络能有效地促进保存系统达到这个目标。

另外在考虑构建WA协作保存网络时，还有一些非常重要的因素本文没有涉及，即构成协作的技术基础，如相应的技术和标准（标准的文件格式、通用的索引和存储、元数据框架、唯一标识符等），本主题的另一篇论文对此有详细介绍。

参考文献

[1] IIPC[EB/OL]. [2009-01-12].http://www.netpreserve.org/about/index.php.
[2] OAIS[EB/OL]. [2009-01-10].http://public.ccsds.org/publications/archive/650x0b1.pdf.
[3] Web archives long term access and interoperability: the International Internet Preservation consortium activity[EB/OL]. [2009-01-10].http://www.ifla.org/IV/ifla71/papers/194e-Lupovici.pdf.
[4] PANDORA[EB/OL]. http://pandora.nla.gov.au/.[2009-02-12].
[5] PANDAS [EB/OL].http://pandora.nla.gov.au/manual/pandas3/workflows.html.[2009-02-12].
[6] PANDORA: Technical Details[EB/OL]. [2009-02-12].http://pandora.nla.gov.au/pandoratech.html.
[7] UKWAC[EB/OL]. [2009-02-12].http://www.webarchive.org.uk/ukwa/.
[8] Chronopolis[EB/OL]. [2009-02-12]. http://chronopolis.sdsc.edu/.
[9] Chronopolis Digital Preservation Framework[EB/OL]. [2009-02-12].http://chronopolis.sdsc.edu/assets/docs/niso_mcdonald.pdf.
[10] NWA.[EB/OL]. [2009-02-12]. http://nwa.nb.no/.
[11] Combine[EB/OL]. [2009-02-12]. http://combine.it.lth.se/.
[12] NEDLIB harvester[R/OL]. [2009-02-12]. http://nedlib.kb.nl/workshop/NEDLIB%20harvester.ppt.
[13] HTTrack[EB/OL]. [2009-03-24].http://www.httrack.com/.
[14] Heritrix[EB/OL]. [2009-03-24]. http://crawler.archive.org/.
[15] Excalibur[EB/OL]. [2009-03-24]. https://excalibur.bnp.org.uk/acatalog/search.html.
[16] FAST search engin[EB/OL]. [2009-03-24].http://www.fastsearch.com/.
[17] Lucene [EB/OL]. [2009-03-24].http://lucene.apache.org/.
[18] NWA framework [EB/OL]. [2009-03-24]. www.ifnet.it/elag2002/ws_paper/ws5.ppt.
[19] The Nordic Web Archive[OB/OL]. [2009-03-24].http://www.deflink.dk/upload/doc_filer/doc_alle/1023_SBA.ppt.
[20] PWA[OB/OL]. [2009-03-24]. http://arquivo-web.fccn.pt/portuguese-web-archive-2?set_language=en.
[21] Introducing the Portuguese web archive initiative[EB/OL]. [2009-03-24].http://arquivo-web.fccn.pt/sobre-o-arquivo/introducing-the-portuguese-web-archive-initiative.
[22] Web At Risk Project[EB/OL]. [2009-03-24]. http://web3.unt.edu/webatrisk/.

作者简介

吴振新，中国科学院国家科学图书馆副研究馆员、研究方向：数字资源长期保存。通讯地址：北京市北四环西路33号、中国科学院国家科学图书馆、100190。E-mail: wuzx@mail.las.ac.cn
张智雄，中国科学院国家科学图书馆研究馆员、研究方向：知识技术。通讯地址：北京市北四环西路33号，中国科学院国家科学图书馆、100190。E-mail: zhangzx@mail.las.ac.cn
王婷，中国科学院国家科学图书馆2006级硕士研究生。通讯地址：北京市北四环西路33号、中国科学院国家科学图书馆、100190。

Research on the Web Archive Cooperative Networks

Wu Zhenxin, Zhang Zhixiong, Wang Ting / National Science Library, Beijing , 100190

Abstract: The cooperative network is using web and software tools to provide cooperative services for people in different physical locations. This paper focuses on the cases about some classical web archive cooperative networks from a technological angle, makes some analysis and research on how to construct cooperative networks to co-share web archive systems and web resources.

Keywords: Web resource, Web archive, Cooperative networks

（收稿日期：2009-05-15；责任编辑：贾延霞）